Doc-Former: A transformer-based document shadow denoising network

Shengchang Pei

Hubei Key Laboratory of Intelligent Robot (Wuhan Institute of Technology) School of Computer Science and Engineering, Wuhan Institute of Technology Wuhan, China

Jun Liu*

Hubei Key Laboratory of Intelligent Robot (Wuhan Institute of Technology) School of Computer Science and Engineering, Wuhan Institute of Technology Wuhan, China

Key Laboratory of Biomarkers and In Vitro Diagnosis Translation of Zhejiang province, China

Niannian Yi

Hubei Key Laboratory of Intelligent Robot (Wuhan Institute of Technology) School of Computer Science and Engineering, Wuhan Institute of Technology Wuhan, China

Yun Zhang

Hubei Key Laboratory of Intelligent Robot (Wuhan Institute of Technology) School of Computer Science and Engineering, Wuhan Institute of Technology Wuhan, China

Zhengtao Liu

Hubei Key Laboratory of Intelligent Robot (Wuhan Institute of Technology) School of Computer Science and Engineering, Wuhan Institute of Technology Wuhan, China

Zengyan Chen

Hubei Key Laboratory of Intelligent Robot (Wuhan Institute of Technology) School of Computer Science and Engineering, Wuhan Institute of Technology Wuhan, China

The existence of shadows makes the visual perception and readability of document images poor, so how to remove the shadows in these document images is an urgent problem to be solved in the industry. Currently, only a few methods are specifically designed for shadow removal of document images. Among them, some algorithms are heuristic algorithms based on experience or direct observation. These algorithms only heuristically denoise the image from the perspective of light or color, and do not consider the specific characteristics of the shadow of the document. So, we propose a transformer-based document shadow denoising algorithm, and the experimental comparison proves that it has achieved state-of-the-art excellence in its performance.

CCS CONCEPTS • Neural Networks • deep learning • transformer • Image Denoising Methods

Keywords: document shadow denoising algorithm; transformer

1 INTRODUCTION

Paper documents are frequently encountered in our daily lives. In modern society, newspapers, receipts, reports, and other printed materials are considered as documents. With the continuous development of technology, we often use the cameras in our smartphones to capture images of documents instead of using scanners. Compared to scanners, the photos taken with a camera appear more like the actual paper documents. The scanned images obtained from scanners or scanning apps often lose the authenticity of the original image. However, there are still issues with using the smartphone camera to capture document images. Due to various factors such as poor lighting conditions, unfavorable angles, or casual photography posture, the document images captured by the camera may inevitably contain shadows. The presence of these shadows greatly affects the visual appearance and legibility of the document images. Therefore, the removal of shadows from document images is a pressing issue in the industry.

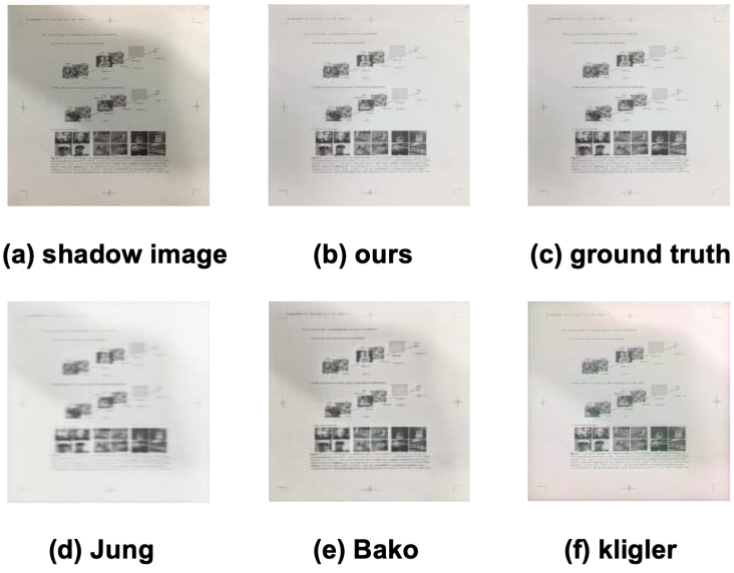

Figure 1: Comparison diagram of several methods focusing on document shadow image denoising.

Currently, only a few methods are specifically designed for shadow removal in document images. Some of these algorithms are based on experience or heuristic approaches derived from observation. For example, Bako et al.[1] proposed a method that analyzes the intensity of the image in shadow and non-shadow regions, and separates the image into two clusters: background and text. They generate a shadow map using the average of the background cluster, effectively removing the shadows. However, as shown in Figure 1 (e), since the authors directly analyze the specific color of the shadow region, when there are variations in the background color or uneven brightness in the background, the resulting shadow removal may introduce artifacts at the boundaries of the shadows, creating some pseudo-shadows in the final output image. Different from Bako et al.'s[1] direct measurement of pixel attributes, Kligler et al.'s[2] shadow removal algorithm directly identifies the most representative 3D points of the visible and occluded parts in document images. Although Kligler's[2] algorithm has shown significant improvement for shadowed document images with rich colors, as shown in

Figure 1(f), some artifacts are still inevitable in images with high shadow intensity. As shown in Figure 1(d), Jung's[3] method achieves excellent shadow removal optimization for document images in complex lighting conditions. However, this algorithm often treats text and photographs with boundaries as shadows and eliminates them, resulting in incomplete generated images. With the continuous development of deep learning, denoising methods such as ST-CGAN[4] and Mask-ShadowGAN[5] have made significant progress in natural image denoising. However, due to the high similarity between document shadows and the dark areas of printed text, many methods still struggle with effectively removing shadows from document images.

In recent years, BEDSR-Net[6], proposed by Yun-Hsuan et al., has emerged as the first network specifically designed for document shadow images, achieving outstanding artistic-level results. The network consists of a background evaluation network and a shadow removal network. It first evaluates the background of the document image and then utilizes a U-net-based[7] generator to generate shadow-free images using the conditional generative adversarial networks (cGANs[8]) loss. BEDSR-Net[6] serves as the benchmark for our main comparison. In this context, we propose the first transformer-based document shadow image denoising network, called doc-former. Unlike the two-stage network of BEDSR-Net[6], we do not require background evaluation or shadow region detection and localization for the document image. Instead, we directly input the noisy original image into the network, which produces an end-to-end output of a shadow-free image. Through experimental comparisons on multiple publicly available datasets, doc-former achieves state-of-the-art performance in the industry. Our main contributions are as follows:

- We propose the first transformer-based document shadow image denoising network, doc-former, which achieves end-to-end document shadow image denoising.
- Our approach innovatively utilizes wavelet transformation and its inverse transformation for lossless downsampling and upsampling of images. This reduces the model size while preserving the performance of the transformer, enhancing model portability, and saving GPU memory consumption.
- We incorporate valuable channel attention in the network, allowing the network to preserve more detailed features.

2 MATERIALS AND METHODS

2.1 The structure of doc-former

The doc-former network that we propose is specifically designed for document shadow image denoising, and it follows a U-shaped symmetric encoder-decoder architecture. As shown in Figure 2, our network structure mainly consists of transformer blocks and CrossAttention blocks. The transformer blocks are responsible for the downsampling and upsampling operations of the image. CrossAttention is utilized to enable the model to differentiate attention between channels and positions. The Discrete Wavelet Transform (DWT) and Inverse Discrete Wavelet Transform (IWT) are employed for lossless downsampling and upsampling of the model.

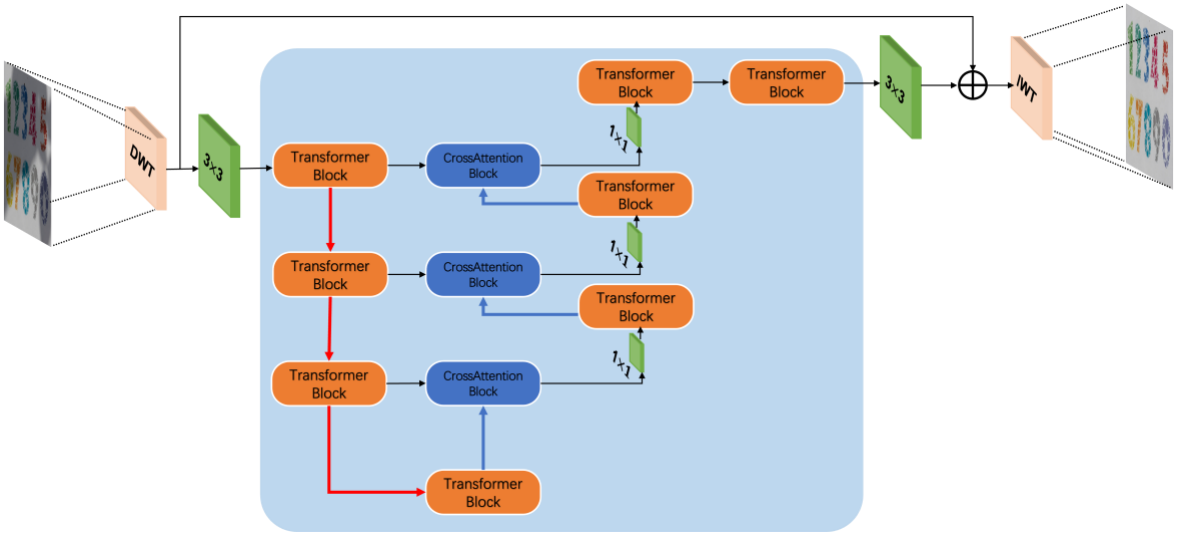

Figure 2: The network structure of doc-former primarily consists of a U-shaped symmetric encoder-decoder architecture.

2.1.1 DWT&IWT

In today's era where transformer models have gained popularity, we enjoy the improved experimental performance they offer. However, we also must deal with the significant GPU memory consumption and resource utilization they bring. Consequently, training transformer-based models becomes challenging on servers with lower GPU performance. Juncheng et al.[9] proposed an efficient wavelet transformer that effectively addresses the issues. Although the source code of their implementation is not currently publicly available, we have been inspired by their work and incorporated the classic Discrete Wavelet Transform (DWT) and Inverse Discrete Wavelet Transform (IWT) at the beginning and end of our network, respectively, to replace one downsampling and upsampling operation. This allows us to achieve similar benefits while mitigating the GPU memory constraints.

$$I_{LL}, I_{LH}, I_{HL}, I_{HH} = f_{DWT}(I_{noisy}) \quad (1)$$

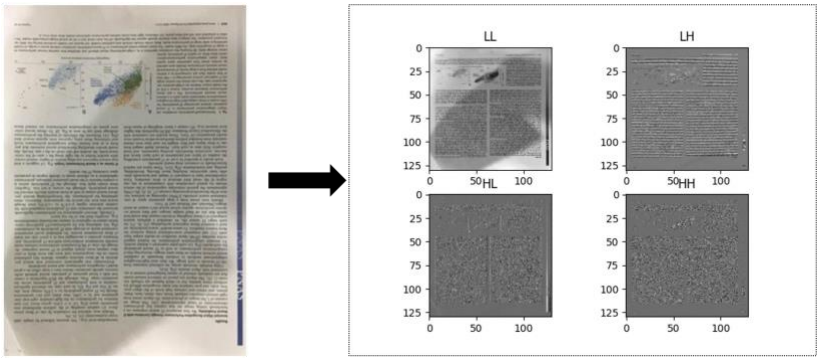

Figure 3: The rendering of the document shadow image after DWT.

2.1.2 Transformer based Encoder-Decoder

In recent years, the development of natural language processing has been rapid, and these huge successes are inseparable from the birth of Transformer. Not only in natural language processing, but some other computer vision tasks have also achieved remarkable results due to Transformer. As Dosovitskiy and others proposed Vision Transformer (ViT[10]), many scholars began to think about applying Transformer's multi-attention mechanism and excellent backbone modules to image feature extraction. For example, U-former[11] and restormer[12] are examples of these applications, and they both show excellent performance in image restoration work. Inspired by restormer[12], we inherited restormer's[12] tansformer block in doc-former. As shown in Figure 4, it consists of a multi-head temporal attention and gated feed-forward network.

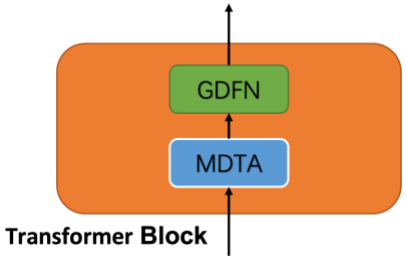

Figure 4: The structure of the transformer block.

2.1.3 Channel attention mechanism

We have designed a novel CrossAttention module as the skip connection between the encoder and decoder, as shown in the diagram. When we designed this module, we did not use a traditional mechanical way to directly connect the features of the encoder layer to the decoder layer. Instead, we adopt a more efficient approach in order to provide spatial information to each decoder, thus helping to generate the output mask and preserve details. As shown in Figure 5, in our architecture, we use skip connections (X2) to pass the output of the encoder layer to the decoder layer, which is used as the input of the query key of the lower layer, and its embedding dimension is $\mathbf{n_2} \times \mathbf{d_q}$. At the same time, we use the output of the lower decoder layer (X1) as input for the keys and values of the next layer, whose embedding dimensions are $\mathbf{n_2} \times \mathbf{d_v}$ and $\mathbf{d_k} \times \mathbf{n_2}$ respectively. It is worth

noting that at the beginning, in order to fuse these two sets of features, we introduce a linear layer to adjust the dimension of X1 to the same embedding dimension $\mathbf{n_2} \times \mathbf{d}$ as X2. We choose to use X2 as query input because this effectively models multi-level representations within the attention block.

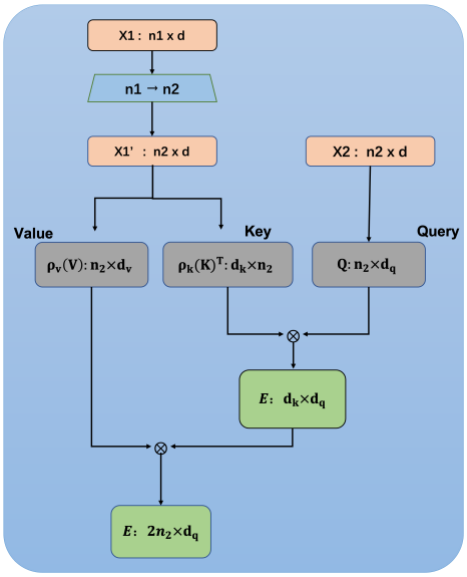

Figure 5: Structure of CrossAttention Block.

$$X_1' = F(X_1) \quad (2)$$

$$K, V = Proj(X_1') \quad (3)$$

$$Q = Proj(X_2) \quad (4)$$

$$E = \rho_V(V)\rho_k(K^T)Q \quad (5)$$

2.1.4 Downsampling&Upsampling

When a document shadow image of size H×W×C is input into the network, after passing through the DWT downsampling layer, the width and height of the image become half of the original size, while the number of channels becomes four times the original. After the 3×3 convolutional layer, the image goes through three transformer blocks, which perform downsampling operations three times. It then undergoes three upsampling operations, with cross-attention blocks used for connection. Afterwards, the refined features from the 3×3 convolutional layer and the shallow features from the DWT are aggregated using a global residual learning strategy to form the final reconstructed features. Finally, the features are upsampled using IWT to restore the document image to its original resolution, resulting in a shadow-free document image.

2.2 Lossfunction

The loss function consists of two main parts: L_GAN and L_1.

L_GAN is a form of loss resistance that is used to train a generator network to produce more realistic results. This method involves a discriminator network that tries to distinguish between generated results and real images (without shading). The goal of the generator is to fool the discriminator by producing results that are indistinguishable from real images. Through this adversarial loss, the generator is encouraged to produce visually convincing shadow-free document images.

$$\mathrm{L}_{GAN}(G,\ D) =$$

$$\varepsilon_{x,y\sim p_{data}(x,y)}[logD(x,y)] + \varepsilon_{x\sim p_{data}(x),\ z\sim p_{z(z)}}[log\ (1 - D(x, G(x,y)))] \quad (6)$$

L_1 loss is the second loss designed when training our model. The function of this part is mainly to evaluate noisy images and noisy images to determine the accuracy and consistency of regenerated pixels after passing through the generator. We will the noisy picture is used as the input I_{input}, and the noise-free picture is used as the output I_{output}. In addition, λ_c represents the weight value of each channel, which rescales the features from different channels. The final ϕ is obtained by the network model through forward propagation and represents the prediction effect of the network. Finally, C, H, and W obviously represent some of the main parameters of the image, which are the number of channels, height value, and width value of the image respectively.

$$L_1(G) = \frac{1}{4HW}\sum_{c=1}^{C}\sum_{v=1}^{H}\sum_{u=1}^{W}\lambda_c|I_{output} - \phi(I_{input})|_1 \quad (7)$$

3 EXPERIMENTS

We compared our method with four state-of-the-art approaches, including traditional heuristic algorithms and the current leading method, BEDSR-Net[6], on the publicly available Jung[3] and Kligler[2] datasets. The experiments were conducted on an A100-PCIE*1 card with 40GB of memory.

The Jung[3] datasets consist of 159 light-distorted document images captured with a smartphone camera. Kligler[2] datasets are composed of 381 artificial shadow images, which contain four categories of shadow images: handwritten documents, printed documents, posters and fonts.

To evaluate the experimental results, we utilized two popular image quality assessment metrics: Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM). PSNR measures the ratio of the peak signal power to the noise power in an image, indicating the level of noise reduction achieved. SSIM assesses the structural similarity between the denoised image and the original image, considering both luminance and contrast information.

Based on the results shown in Figure 6, our method outperforms the current state-of-the-art document shadow image denoising algorithm, BEDSR-Net[6]. Our method demonstrates superior performance in handling document images with highly complex lighting conditions.

The figure provides visual evidence of the effectiveness of our method in removing shadows and improving the quality of document images. It shows a comparison between the output of our method and the output of BEDSR-Net[6] on challenging document shadow images. The improved image quality achieved by our method can be attributed to the innovative techniques and network architecture specifically designed for document shadow image denoising.

Not only that, but we also conduct ablation experiments on our method. We remove channel attention from our network, and experimental results demonstrate that the denoising performance is affected after removal. However, since our network uses a transformer-based symmetric network, the network after removing channel attention still surpasses the existing algorithms.

These results highlight the capability of our method to handle diverse lighting scenarios and successfully reduce the impact of shadows, ultimately enhancing the overall quality and usability of document images.

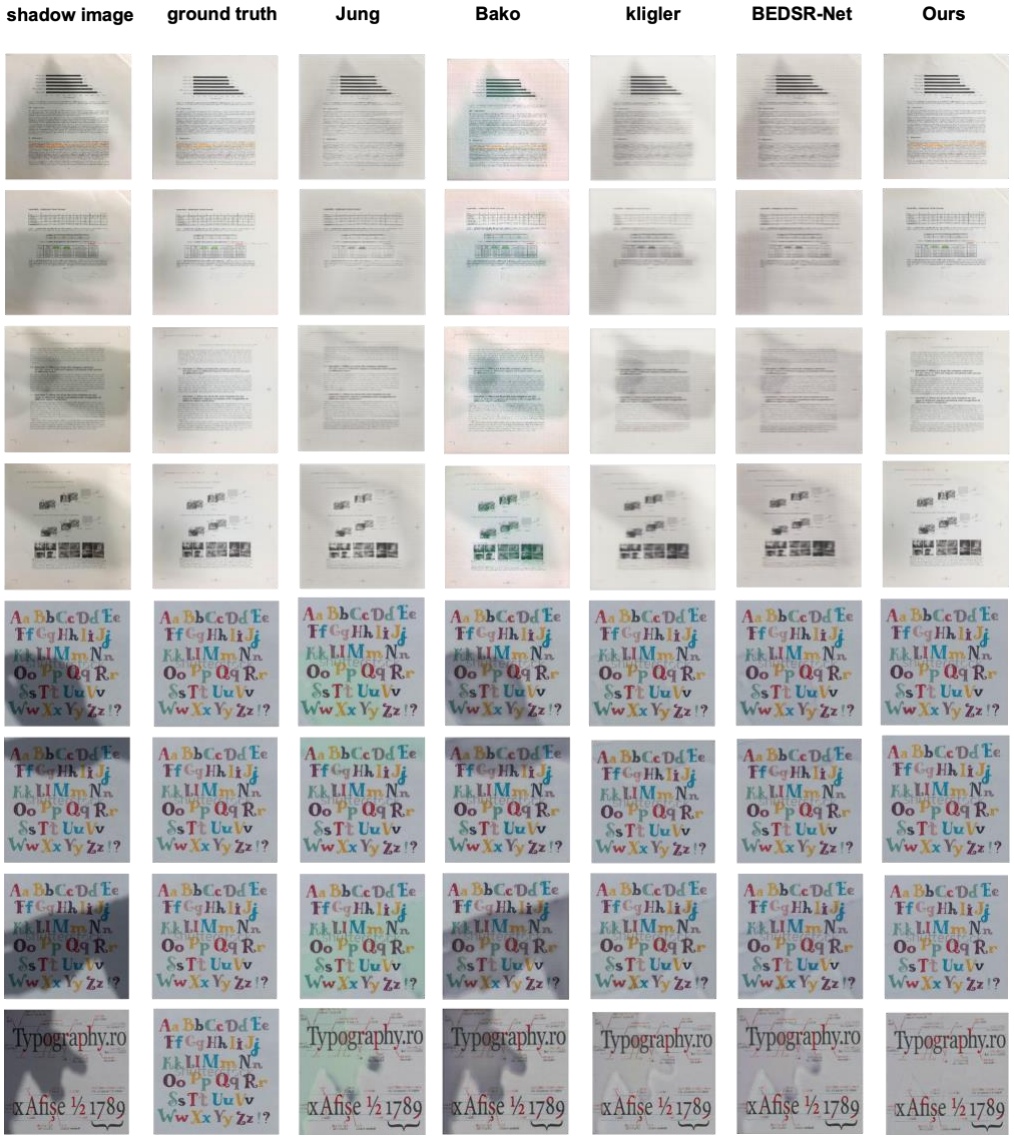

Figure 6: Select 4 photos on the Jung[3] and Kligler[2] datasets to compare the five methods.

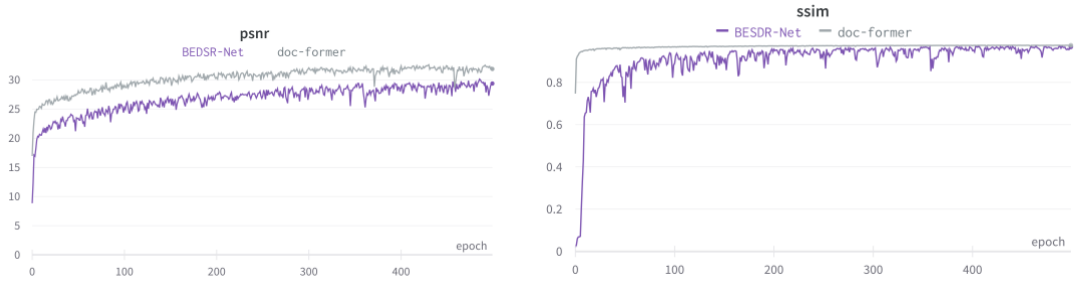

Figure 7: Comparison of our method and BEDSR-Net[6] training 500 rounds on the Jung[3] dataset.

Table 1: Prediction results under different wavelength screening methods.

Dataset	Metrics	Jung[3]	Kligler[2]	Bako[1]	BEDSR-Net[6]	Doc-Former w/o CA block	Ours
Jung[3]'s dataset	PSNR	28.24	24.56	22.43	28.64	30.21	33.96
	SSIM	0.9219	0.8417	0.9034	0.9217	0.9533	0.9786
Kligler's[2] dataset	PSNR	19.30	26.67	29.17	30.19	32.43	35.23
	SSIM	0.8534	0.8536	0.9153	0.9320	0.9650	0.9820

4 CONCLUSION

Our proposed method, doc-former, offers a novel approach in the field of document shadow image denoising. Experimental results demonstrate that doc-former, based on the transformer architecture, outperforms traditional heuristic algorithms and establishes a notable advantage over the state-of-the-art CNN-based BEDSR-Net[6].

ACKNOWLEDGMENTS

This work was supported by Industry-University-Research Innovation Fund of Science and Technology Development Center, MOE[2022TX004], the Educational Research Foundation of Chinese Society of Academic Degrees and Graduate Education [2020MSB2], the Industry-University Cooperation and Collaborative Education Project of the Ministry of Education [202102564003], Science and Technology Development Funds of the State Administration of Market Regulation [2021MK071], the Natural Science Foundation of Hubei Province [2022CFC001], the Foundation of Hubei Provincial Key Laboratory of Intelligent Robot [HBIRL 202209], the Opening Fund of Key Laboratory of Biomarkers and In Vitro Diagnosis Translation of Zhejiang province [KFJJ 2023006], the Fourteenth Graduate Innovation Fund of Wuhan Institute of Technology [CX2022331,CX2022348,CX2022365].

REFERENCES

[1] Steve Bako, Soheil Darabi, Eli Shechtman, Jue Wang, Kalyan Sunkavalli, and Pradeep Sen. 2016. Removing shadows from images of documents. In Proceedings of Asian Confer- ence on Computer Vision (ACCV), pages 173–183.

[2] Netanel Kligler, Sagi Katz, and Ayellet Tal. 2018. Document en- hancement using visibility detection. In Proceedings of the IEEE

Conference on Computer Vision and Pattern Recogni- tion (CVPR), pages 2374–2382.

[3] Seungjun Jung, Muhammad Abul Hasan, and Changick Kim. 2018. Water-filling: An efficient algorithm for digitized doc- ument shadow removal. In Proceedings of Asian Conference on Computer Vision (ACCV), pages 398–414.

[4] Jifeng Wang, Xiang Li, and Jian Yang. 2018. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1788–1797.

[5] Shengfeng He, Bing Peng, Junyu Dong, and Yong Du. 2021. Mask-shadownet: Toward shadow removal via masked adaptive instance normalization. IEEE Signal Process- ing Letters, vol. 28, pp. 957–961.

[6] Y. -H. Lin, W. -C. Chen and Y. -Y. Chuang. 2020. BEDSR-Net: A Deep Shadow Removal Network From a Single Document Image. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 12902-12911. doi: 10.1109/CVPR42600.2020.01292.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U- Net: Convolutional networks for biomedical image segmen- tation. In Proceedings of International Conference on Med- ical image Computing and Computer-Assisted Intervention (MICCAI), pages 234–241. Springer.

[8] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

[9] Li, J., Cheng, B., Chen, Y., Gao, G., & Zeng, T. 2023. EWT: Efficient Wavelet-Transformer for Single Image Denoising. ArXiv, abs/2304.06274.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby .2021. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR.

[11] ZhendongWang,XiaodongCun,JianminBao,andJianzhuang Liu. 2021. Uformer: A general u-shaped transformer for image restoration. arXiv preprint 2106.03106.

[12] Zamir S W, Arora A, Khan S, et al. 2022. Restormer: Efficient transformer for high-resolution image restoration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5728- 5739.

Authors' background

Your Name	**Title***	**Research Field**	**Personal website**
Shengchang Pei	master student	Machine Learning	shengchangpei@stu.wit.edu.cn
Jun Liu*	full professor	Machine Learning	liujun@wit.edu.cn
Niannian Yi	master student	Machine Learning	347048558@qq.com
Yun Zhang	master student	Machine Learning	keys.zw@qq.com
Zhengtao Liu	master student	Machine Learning	2295079374@qq.com
Zengyan Chen	master student	Machine Learning	2423574809@qq.com